KYOTO: A System for Mining, Structuring, and Distributing Knowledge Across Languages and Cultures

Piek Vossen[1, 2], Eneko Agirre[3], Nicoletta Calzolari[4], Christiane Fellbaum[5, 6], Shu-kai Hsieh[7], Chu-Ren Huang[8], Hitoshi Isahara[9], Kyoko Kanzaki[9], Andrea Marchetti[10], Monica Monachini[4], Federico Neri[11], Remo Raffaelli[11], German Rigau[3], Maurizio Tesconi[10], Joop VanGent[2]

[1] Faculteit der Letteren, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081HV Amsterdam, Netherlands, p.vossen@let.vu.nl
[2] Irion technologies, Delftechpark 26, 2628XH Delft, Netherlands {piek.vossen, gent}@irion.nl
[3] IXA NLP group, University of the Basque Country, Manuel Lardizabal 1, Donostia, Basque -Country {e.agirre, g.rigau@ehu.es}
[4] Istituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche, Via Moruzzi 1, 56124 Pisa, Italy, nicoletta.calzolari@ilc.cnr.it, monica.monachini@ilc.cnr.it,
[5] Berlin-Brandenburg Academy of Sciences, Berlin, Germany
[6] Princeton University, Princeton, USA
[7] National Taiwan Normal University, Republic of China
[8] Academia Sinica, Taipei, Republic of China
[9] NICT, Kyoto, Japan
[10] Istituto di Informatica e Telematica, Consiglio Nazionale delle Ricerche, Via Moruzzi 1, 56124 Pisa, Italy, andrea.marchetti@iit.cnr.it, maurizio.tesconi@iit.cnr.it
[11] Synthema, Pisa, Italy

Abstract. We outline work to be carried out within the framework of an impending EC project. The goal is to construct a language-independent information system for a specific domain (environment/ecology) anchored in a language-independent ontology that is linked to wordnets in several languages. For each language, information extraction and identification of lexicalized concepts with ontological entries will be done by text miners ("Kybots"). The mapping of language-specific lexemes to the ontology allows for crosslinguistic identification and translation of equivalent terms. The infrastructure developed within this project will enable long-range knowledge sharing and transfer to many languages and cultures, addressing the need for global and uniform transition of knowledge beyond the domain of ecology and environment addressed here.

Keywords: Global Wordnet Grid, ontologies and wordnets, multilinguality, semantic indexing and search, text mining.

1 Introduction

Economic globalization brings challenges and the need for new solutions that can serve all countries. Timely examples are environmental issues related to rapid growth and economic developments such as global warming. The universality of these problems and the search for solutions require that information and communication be supported across a wide range of languages and cultures. Specifically, a system is needed that can gather and represent in a uniform way distributed information that is structured and expressed differently across languages. Such a system should furthermore allow both experts and laymen to access information in their own language and without recourse to cultural background knowledge.

Addressing sudden and unpredictable environmental disasters (fires, floods, epidemics, etc.) requires immediate decisions and actions relying on information that may not be available locally. Moreover, the sharing and transfer of knowledge are essential for sustainable growth and long-term development. In both cases, it is important that information and experience are not only distributed to assist with local emergencies but are universally re-usable. In these settings, natural language is the most ubiquitous and flexible interface between users -- especially non-experts -- and information systems.

The goal of "**K**nowledge-**Y**ielding **O**ntologies for **T**ransition-Based **O**rganization" (KYOTO) is, first, to develop a content enabling system that provides deep semantic search. KYOTO will cover access to a broad range of multimedia data from a large number of sources in a variety of culturally diverse languages. The data will be accessible to both experts and the general public on a global scale.

KYOTO is funded under project number 211423 in the 7th Frame work project in the area of Digital Libraries: FP7-ICT-2007-1, Objective ICT-2007.4.2: Intelligent Content and Semantics. It will start early 2008 and last 3 years. The consortium consists of research institutes, companies and environmental organizations: Vrije Universiteit Amsterdam (Amsterdam, The Netherlands), Consiglio Nazionale delle Ricerche (Pisa, Italy), Berlin-Brandenburg Academy of Sciences and Humantities (Berlin, Germany), University of the Basque Country (Donostia, Basque Country), Academia Sinica (Tapei, Taiwan), National Institute of Information and Communications Technology (Kyoto, japan), Irion Technologies (Delft, The Netherlands), Synthema (Pisa, Italy), European Centre for Nature Conservation (Tilburg, The Netherlands), World Wide Fund for Nature (Zeist, The Netherlands), Masaryk University (Brno, Czech). In total 364 person months of work are involved. The partners from Taiwan and Japan are funded by national grants.

2 The KYOTO System: Overview

KYOTO is a generic system offering knowledge transition and information across different target groups, transgressing linguistic, cultural and geographic boundaries. Initially developed for the environmental domain, KYOTO will be usable in any knowledge domain for mining, organizing, and distributing information on a global scale in both European and non-European languages.

KYOTO's principle components are an ontology linked to wordnets in a broad range of languages (Basque, Chinese, Dutch, English, Italian, Japanese, Spanish), linguistic text miners, a Wiki environment for supporting and maintaining the system, and a portal for the environment domain that allows for deep semantic searches.

Concept extraction and data mining are applied through a chain of semantic processors ("Kybots") that share a common ground and knowledge base and re-use the knowledge for different languages and for particular domains.

Information access is provided through a cross-lingual user-friendly interface that allows for high-precision search and information dialogues for a variety of data from wide-spread sources in a range of different languages. This is made possible through a customizable, shared ontology that is linked to various wordnets and that guarantees a uniform interpretation for diverse types of information from different sources and languages.

The system can be maintained and kept up to date by specialists in the field using an open Wiki platform for ontology maintenance and wordnet extension.

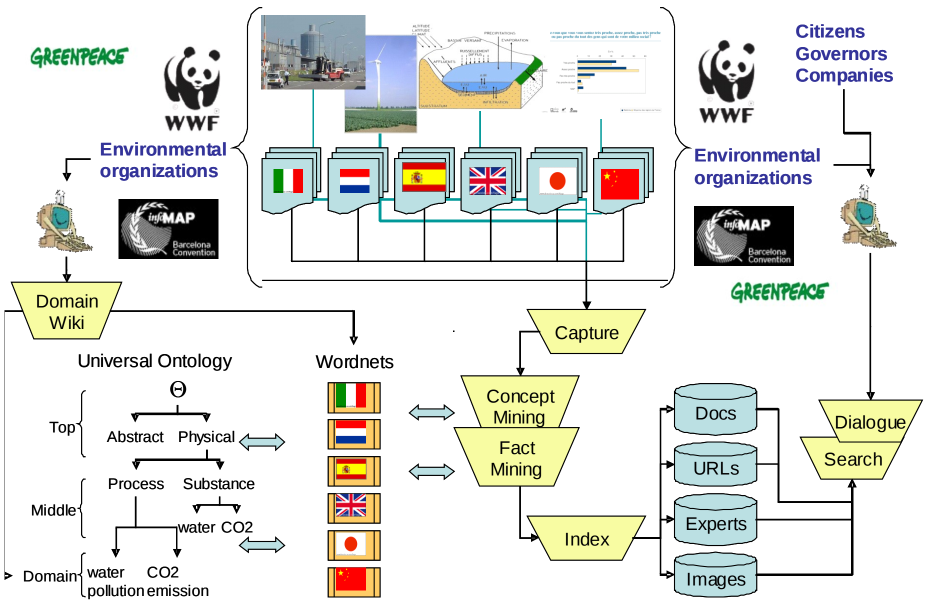

Fig. 1. System architecture

Figure 1 gives an overview of the complete system. In this schema, information stored in various media and languages, distributed over different locations, is collected through a **Capture** module and stored in a uniform XML representation. For each language, **concept miners** are applied to derive concepts that occur in the textual data and compare these with the given wordnets for the different languages. The wordnets provide a mapping to a single shared **ontology**. Both the wordnets and the ontology can be modified and edited in a special **Wiki** environment by the people in a community; in the present project, these will be specialists in the environment domain. Encoding of knowledge and wordnets for a domain will result in more precise and effective mining of information and data through **fact mining** by the so-called **Kybots**. Kybots will be able to detect specific patterns and relations in text because of the concepts and constraints coded by the experts. These relations are added to the XML representation of the captured text. An **indexing module** then creates the indexes for different databases and data types that can be accessed by the users through a text **search** interface or possibly **dialogue** systems. The users can be the same environmental organisations, and/or governments and citizens.

In the next sections, we will discuss KYOTO's major components in more detail.

3 The Ontology

The ontology, where knowledge of concepts is formally encoded, consists of three layers. The top layer is based on existing top level ontologies, among them SUMO [10, 11], DOLCE [8] and the MEANING Top Concept Ontology [3]. We will investigate what ontology will be the best basis for our purpose and can also be shared across the diverse languages and cultures. If necessary, ontology fragments or elements can be shared or a selection will be made. We do not expect major differences in the fundamental semantic organisation of the different languages. Recent studies, for example, show that the Chinese radical system and character compounding tend to be based on the same qualia distinctions as in the Generative Lexicon [4, 5].

The middle layer will be derived from existing wordnets, where concepts are mapped to lexical units. The ontology's mid-level must be developed such that it connects domain terms and concepts to the top-level. We define all the high-level and mid-level concepts that are needed to accommodate the information in the environmental domain. Knowledge is implemented at the most generic level to maximize re-usability yet precisely enough to yield useful constraints in detecting relations. Within the domain, we extend the ontologies to cover all necessary concepts and applicable, sharable relations.

The domain terms are extracted semi-automatically from the source documents or manually created through a Domain Wiki. The Domain Wiki allows experts to modify and extend the domain level of the ontology and extend the wordnets accordingly. It enables community-based resource building, which will lead to increased, shared understanding of the domain and at the same time result in the formalization of this knowledge, so that it can be used by an automatic system.

This resource will build on the Multilingual Central Repository (MCR) knowledge base [1] developed in the MEANING project [12]. Currently, the MCR consistently integrates more than 1.6 million semantics links among concepts. Moreover, the current MCR has been enriched with about 460.000 semantic and ontological properties [2]: Base Concepts and Top Concept Ontology [3], WordNet Domains [7], Suggested Upper Merged Ontology (SUMO) [10], providing ontological coherence to all the uploaded wordnets.

Extensions to wordnets and the ontology will be propagated through appropriate sharing protocols, developed exploiting LeXFlow, a framework for rapid prototyping of cooperative applications for managing lexical resources (XFlow [9] and LeXFlow

[13, 14, 15, 16]). The shared ontology guarantees a uniform interpretation layer for the diverse information from different sources and languages. At the lowest level of the ontology, we expect that abstract constraints and structures can be hidden for the users but can still be used to prevent fundamental errors, e.g. creating a concrete concept for an adjective. The Wiki users should focus on formulating conditions and specifications that they understand without having to worry about the linguistic and knowledge engineering aspects. They can discuss these specifications within their community to reach consensus and provide proper labels in each language.

4 Kybots

Once the ontological anchoring is established, it will be possible to build text mining software that is able to detect semantic relations and propositions. Data miners, so-called Kybots (**K**nowledge-**y**ielding ro**bots**), can be defined using constraints among relations at a generic ontological level. These logical expressions need to be implemented in each language by mapping the conceptual constraint onto linguistic patterns. A collection of Kybots created in this way can be used to extract the relevant knowledge from textual sources represented in a variety of media and genres and across different languages and cultures. Kybots will represent such knowledge in a uniform and standardized XML format, compatible with WWW specifications for knowledge representation such as RDF and OWL.

Kybots will be developed to cover users' questions and answers as well as generic concepts and relations occurring in any domain, such as named-entities, locations, time-points, etc. Kybots are primarily defined at a generic level to maximize re-usability and inter-operability. We develop the Kybots that are necessary for the selected domain but the system can easily be extended and ported to other domains.

The Kybots will operate on a morpho-syntactic and semantic encoding level that will be the same across all the languages. Every group will use existing linguistic processors or develop additional ones when needed to foresee in a basic linguistic analysis, which involves: tokenization, segmentation, morpho-syntactic tagging, lemmatization and basic syntactic parsing. Each of these processes can be different but the XML encoding of the output will be the same. This will guarantee that Kybots can be applied to the output of text in different languages in a uniform way. We will use as much as possible existing and available free software for this process. Note that the linguistic expression rules of ontological patterns in a specific Kybot are to be defined on the basis of the common output encoding of the linguistic processors. Likewise, they can share specifications of linguistic expression in so far the relations are expressed in the same way in these languages.

5 Indexing, Searching, and Interfacing

The extracted knowledge and information is indexed by an existing search system that can handle fast semantic search across languages. It uses so-called contextual conceptual indexes, which means that occurrences of concepts in text are interpreted by their co-occurrence with other concepts within a linguistically defined context, such as a noun phrase or sentence. The co-occurrence patterns of concepts can be specified in various ways, possibly based on semantic relations that are defined in the logical expressions. Thus, the system yields different results for searches for *polluting substance* and *polluted substance*, because these involve different semantic relations between the same concepts. By mapping a query to concepts and relations, very precise matches can be generated, without the loss of scalability and robustness found in regular search engines that rely on string matching and context windows.

Reasoning over facts and ontological structures will make it possible to handle diverse and more complex types of questions. Crosslinguistic and crosscultural understanding is vouchsafed through the ontological anchoring of language via wordnets and text miners.

6 The Wiki Environment

The Wiki environment enables domain experts to easily extend and manage the ontology and the wordnets in a distributed context, to constantly reflect the continuous growth and changes of the data they describe. It owns the characteristics typical of a generic Wiki engine:

- Web based highly-interactive interface, tailored to domain experts who don't know the underlying complex data model (ontology plus wordnet of different languages);
- tools to support collaborative editing and consensus achievement such as discussion forums, and list of last updates;
- automatic acquisition of information from external Web resources (e.g. Wikipedia);
- rollback mechanism: each change to the content is versioned;
- search functions providing the possibility to define different search patterns (synset search, textual search and so on);
- role-based user management.

In addition, the Wiki engine manages the underlying complex data model of the ontology and the wordnets so as to keep it consistent: this is achieved through the definition of appropriate sharing protocols. For instance, when a new domain term such as *water pollution* is inserted into a language-specific wordnet by a domain expert, a new entry, referred to as dummy entry because of the incompleteness of the information represented, will be automatically created and added to the ontology and in the remaining wordnets. The Wiki environment will list all dummy entries to be filled in, in order to notify them to domain experts allowing for their complete definition and integration into KYOTO ontological and lexical resources. In this context, English can be used as the common ground language in order to support the extension process and the propagation of changes among the different wordnets and the ontology.

7 Sharing

Knowledge sharing is a central aspect of the KYOTO system and occurs on multiple levels.

7.1 Sharing and Re-Use of Generic Knowledge

Sharing of generic ontological knowledge in the domain takes place mainly through subclass relations. We collect all the relevant terms in each language for the domain and add them to the general ontology. Possibly, these concepts can be imported from a specific wordnet and "ontologized." It will be important to specify exactly the ontological status of the terms. Only disjunct types need to be added [6]. For example, *CO2* is a type of substance, whereas *greenhouse gases* do not represent a different type of gas or substance but refers to substances that play a specific role in specific circumstances. In so far as new definitions and axioms need to be specified, they can be added for the specific subtypes in the domain. However, this is only necessary if the related information also needs to be mined from the text and is not already covered by the generic miners. Next, the generic and domain knowledge is shared among all participating languages through the mapping of the different wordnets to the ontology.

Extension to different domains is possible though not within the scope of the current project.

7.2 Sharing and Re-Use of Generic Kybots

The sharing of Kybots is more subtle. For example, *concentrations of substances*, *causal relations between processes* or *conditional states for processes* can be stated as general conceptual patterns using a simple logical expression. Within a specific domain, any of these relations and conditions could be detected in the textual data by just using these general patterns. Thus, people usually do not use special words in a language to refer to the causal relation itself but they use general words such as "cause" or "factor". Since any causal relation may hold among processes and or states, they can also hold in the environmental domain. Certain valid conditions can be specified in addition to the general ones, as they are relevant for the users. For example, CO_2 emissions can be derived from a certain process involving certain amounts of the substance CO_2 but critical levels can be defined in the text miner as a conceptual constraint. Furthermore, we may want to limit the ambiguity of interpretation that arises at the generic levels to only one interpretation at the domain level; it is currently an open question to what extent generic patterns can be used or need to be tuned.

Each language group can build a Kybot, capturing a particular relation. A given logical expression that underlies the Kybot of another language can be re-used, or a new pattern can be formulated for a language and a generic universal pattern derived from it. We foresee a system where the text miners can load any set of Kybots in combination with the ontology, a set of wordnets and expression rules in each language. Each Kybot, a textual XML file, contains a logical expression with constraints from the ontology (either the general ontology or a domain instantiation). Through the wordnets and the expression rules, the text miner knows how to detect a pattern in running text for each specific language. In this way, logical patterns can be shared across languages and across domains.

A Kybot can likewise be developed by a group in one language and taken up by another group to apply it to another language. Consider the case where a generic linguistic text miner is formulated for Dutch, based on Dutch words and expressions. This Kybot is projected to the ontology via the Dutch wordnet, becoming a generic ontological expression which relates two ontological classes: a Substance to a Process. This expression may be extended to a domain, where it is applied to CO2 and CO2 emissions. Next, the Spanish group can load the domain specific expression and transform it into a Spanish Kybot that can be applied to a domain text in Spanish. To turn an ontological expression into a Kybot, language expressions rules and functions need to be provided. This process can be applied to all the participating languages, where the basic knowledge is shared.

7.3 Cross-Linguistic Sharing of Ontologies

KYOTO will thus generate Kybots in each language that go back to a shared ontology and shared logical expressions. Thus, KYOTO can be seen as a sophisticated platform for anchoring and grounding meaning within a social community, where meaning is expressed and conceived differently across many languages and cultures. It also immediately makes this shared knowledge operational so that factual knowledge can be mined from unstructured text in domains. KYOTO supports interoperability and sharing across these communities since much knowledge can be re-used in any other domain, and the ontologies support both generic and domain-specific knowledge.

8 Evaluation

The KYOTO system is evaluated in various ways:

1. Wordnets and ontologies are evaluated across linguistic partners;
2. Language and ontology experts will use the Wiki system to build the basic ontology and wordnet layers needed for the extension to the domain;
3. Domain experts will use the top layer and middle layer of wordnets and ontologies plus the Wiki system to encode the knowledge in their domains and reach consensus;
4. The system is tested by integration in a retrieval system;

Cross-linguistic re-use and agreement on the semantic organization is the prime evaluation of the architecture and the system. Proposals for concepts are verified by other wordnet builders and need to be agreed across the languages and cultures. The same happens by domain experts in their domain, except that they do not need to discuss the technical conceptual issues. Both groups will extensively use the Wiki environment to reach agreements and consensus.

The application driven evaluation will use a baseline evaluation that uses the current indexing and retrieval system and the multilingual wordnet database. The knowledge in KYOTO will lead to more advanced indexes in those cases that Kybots have been able to detect the relations in the text. These will lead to more precision in the indexes and also make it possible to detect complex queries for these relations. The performance if the system will be evaluated with respect to the baseline systems. This will be done in two ways:

1. using an overall benchmark system that runs a fixed set of queries on the different indexes and compares the results;
2. using end-user scenarios and interviews carried out on different indexes by test persons;

The questions and queries are selected to show the capabilities of deep semantic processing. They will be harvested from current portals in the environmental domain.

Finally, we plan to give public access to the databases (ontologies and wordnets) and to the retrieval system through the project website. Visitors are invited to try the system and give feedback.

9 Summary and Outlook

KYOTO will represent a unique platform for knowledge sharing across languages and cultures that can represent a strong content based standardisation for the future that enables world wide communication.

KYOTO will advance the state-of-the-art in semantic processing because it is a unique collaboration that bridges technologies across semantic web technologies, wordnet development and acquisition, data and knowledge mining and information retrieval.

On top of the systems and data described earlier, we will build a Wiki environment that will allow communities to maintain the knowledge and information, without expert knowledge of ontologies, knowledge engineering and language technology. The system can be used by other groups and for other domains. Through simple and clear interfaces that exploit the generic knowledge and check the underlying structures, users can reach semantic agreement on the definition and interpretation of crucial notion in their domain. The agreed knowledge can be taken up by generic Kybots that can then detect possible relations on the basis of this knowledge in text that will be indexed and made searchable. All knowledge resources in KYOTO will be public and open source (GPL). This applies to the ontology and the wordnets mapped to the ontology. The GPL condition also applies to the data miners in each language, the DEB servers, the LeXFlow API and the Wiki environments. Any research group should be able to further develop the system, to integrate their own language and/or to apply it to any other domain.

Acknowledgement

The work described here is funded by the European Community, 7th Framework

References

1. Atserias J., L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, P. Vossen. 2004. *The MEANING Multilingual Central Repository*. In Proceedings of the Second International WordNet Conference-GWC 2004 p. 23-30 January 2004, Brno, Czech Republic. ISBN 80-210-3302-9
2. Atserias J., Climent S., Rigau G. *Towards the MEANING Top Ontology: Sources of Ontological Meaning.* LREC'04. ISBN 2-9517408-1-6. Lisboa, 2004.
3. Atserias J., Climent S., Moré J. and Rigau G. 2005. *A proposal for a Shallow Ontologization of WordNet*. Proceedings of the 21th Annual Meeting of the Sociedad Española para el Procesamiento del Lenguaje Natural, SEPLN'05. Granada, España. Procesamiento del Lenguaje Natural num. 35 pp. 161-167. ISSN: 1135-5948.
4. Chou, Ya-Min and Chu-Ren Huang. 2006. *Hantology - A Linguistic Resource for Chinese Language Processing and Studying.* In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy.
5. Chou, Ya-Min, Shu-Kai Hsieh and Chu-Ren Huang. 2007. *Hanzi Grid: Toward a Knowledge Infrastructure for Chinese Character-Based Cultures.* In: Ishida, T., Fussell, S.R., Vossen, P.T.J.M. eds.: *Intercultural Collaboration I. Lecture Notes in Computer Science,* Springer-Verlag
6. Fellbaum, C. and Vossen. P. (2007). ``Connecting the Universal to the Specific: Towards the Global Grid." In: Proceedings of the First International Workshop on Intercultural Communication. Reprinted in: ``Intercultural Collaboration: First International Workshop." Lecture Notes in Computer Science , Vol. 4568, eds. Ishida, T. Fussell, S. R. and Vossen, P.T. J. M. New York: Springer, pp. 1-16.
7. Magnini B., G. Cavaglia. 2000. *Integrating Subject Field Codes into WordNet.* In Gavrilidou M., Carayannis G., Markantonatu S., Piperidis S. and Stainhaouer G. (Eds.) Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation, Athens, Greece, 31 May- 2 June 2000, pp. 1413-1418
8. Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari A. (2003) WonderWeb Deliverable D18 Ontology Library, IST Project 2001-33052 WonderWeb: Ontology Infrastructure for the Semantic Web Laboratory For Applied Ontology - ISTC-CNR, Trento.
9. Marchetti, M. Tesconi, F. Ronzano, M. Rosella, F. Bertagna, M. Monachini, C. Soria, N. Calzolari, C.R. Huang, S.K. Hsieh, 2006. "Towards an Architecture for the Global Wordnet Initiative". In Proceedings of the 3rd Italian Semantic Web Workshop Semantic Web Applications and Perspectives (SWAP 2006), Pisa, Italy 18-20 December, 2006.
10. Niles, I., and Pease, A. 2001. Towards a Standard Upper Ontology. In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Chris Welty and Barry Smith, eds, Ogunquit, Maine, October 17-19, 2001.
11. Pease, A. 2003. The Sigma Ontology Development Environment, in Working Notes of the IJCAI-2003 Workshop on Ontology and Distributed Systems. Proceedings of CEUR 71.
12. Rigau G., Magnini B., Agirre E., Vossen P. and Carroll J. 2002. *MEANING: A Roadmap to Knowledge Technologies.* Proceedings of COLING Workshop "A Roadmap for Computational Linguistics". Taipei, Taiwan. 2002.
13. Tesconi, M. A. Marchetti, F. Bertagna, M. Monachini, C. Soria, N. Calzolari. 2006. "LeXFlow: a framework for cross-fertilization of computational lexicons". In Proceedings of COLING/ACL 2006 Interactive Presentation Session, 17-21 July 2006 Sydney, Australia.

14. Tesconi, M., A. Marchetti, F. Bertagna, M. Monachini, C. Soria, N. Calzolari. 2007. LeXFlow: a Prototype Supporting Collaborative Lexicon Development and Cross-fertilization. In Intercultural Collaboration, First International Workshop, IWIC 2007, Demo and Poster session, Kyoto, Japan
15. Soria, C., M. Tesconi, F. Bertagna, N. Calzolari, A. Marchetti, M. Monachini. 2006. "Moving to dynamic computational lexicons with LeXFlow". Proceedings LREC2006 22-28 May 2006, Genova, Italy.
16. Soria, C., M. Tesconi, A. Marchetti, F. Bertagna, M. Monachini, C.R. Huang and N. Calzolari 2006. "Towards agent-based cross-lingual interoperability of distributed lexical resources". In Proceedings of COLING-ACL Workshop on Multilingual Lexical Resources and Interoperability, 22-23 July 2006, Sydney, Australia.

Relevant URLs

XML and Database Ronald Bourret:	http://www.rpbourret.com/index.htm
Wiki engines:	http://c2.com/cgi/wiki?WikiEngines
Global WordNet Association:	http://www.globalwordnet.org
Princeton WordNet:	http://wordnet@cogsci.princeton.edu
Chinese WordNet:	http://bow.sinica.edu.tw
Italian WordNet:	http://www.ilc.cnr.it/iwndb_php/
LeXFlow:	http://xmlgroup.iit.cnr.it:8888/xflow/login
Multilingual WordNet Service:	http://xmlgroup.iit.cnr.it:88/exist/wordnet/wordnet/
MCR:	http://adimen.si.ehu.es/cgi-bin/wei5/public/wei.consult.perl